

Advanced Natural Language Processing Techniques to Profile Cybercriminals

PROYECTO DE GRADO

AUTOR Alejandro ANZOLA ÁVILA

DIRECTOR Daniel Orlando DÍAZ LÓPEZ, *PhD*

20 de mayo de 2019

Escuela Colombiana de Ingeniería Julio Garavito

Agenda

1. Objetivos y justificación

2. Resultados propuestos y productos obtenidos

3. Marco teórico

Clasificador NAÏVE BAYES

Clasificación con SUPPORT VECTOR MACHINES (SVM)

SELF-ORGANIZING MAPS (SOM)

4. Problemas y soluciones

MODELO 1: Predicción de etiquetas de Twitter

MODELO 2: Reconocimiento de NAMED ENTITIES con redes LSTM

5. Conclusiones y trabajo futuro

Objetivos y justificación

Objetivo general

Generar herramientas y estrategias para el perfilado de cibercriminales con ayuda de metodologías de *NLP* aplicado a datos recolectados de comunicaciones y redes sociales.

Objetivos específicos

- Diseñar e implementar una solución de lenguaje natural para realizar el perfilado de sospechosos.
- Identificar el estado del arte en sistemas que usan *NLP* para apoyar agencias de seguridad del Estado.
- Implementación de artefactos para la construcción de *datasets* con información recolectada de medios privados como de fuentes abiertas.
- Validar la solución desarrollada frente a un escenario real.
- Modelado de diferentes metodologías, heurísticas y meta–heurísticas para *NLP*.

Por hacer

Resultados propuestos y productos obtenidos

Por hacer

1. Entendimiento de las generalidades de DATA SCIENCE:
 - Tipos de MACHINE LEARNING
 - Sistemas de detección de anomalías
 - Diferentes modalidades de clustering
2. Identificación de modelos de NLP aplicables para el perfilado de cibercriminales
3. Entendimiento de los modelos de clasificación y clustering:
 - Clasificador de Naïve Bayes
 - Maquinas de soporte vectorial
 - Mapas autoorganizados
4. Entendimiento de los modelos utilizados en NLP:
 - Predicción de etiquetas con modelos de regresión lineal
 - Reconocimiento de NAMED ENTITIES

- Uso de *embeddings* generados con STARSPACE para los k textos mas similares
5. Implementación de modelos de NLP para el perfilado de cibercriminales:
- Modelo de predicción de hashtags de Twitter con modelos lineales
 - Modelo de reconocimiento de NAMED ENTITIES con redes LSTM
 - Modelo de clustering en redes SOM con *embeddings* de STARSPACE

Marco teórico

Para variables aleatorias x e y , se tiene que la probabilidad condicional $P(y | x)$ es definida como

$$P(y | x) = \frac{P(x | y)P(y)}{P(x)}$$

Clasificador NAÏVE BAYES

Un clasificador de Naïve Bayes estima la probabilidad condicional de las clases por medio de suponer que los atributos son condicionalmente independientes, dado la etiqueta de clasificación y . Donde cada conjunto de d atributos $\mathbb{X} = \{x_1, \dots, x_d\}$ se tiene

$$P(\mathbb{X} \mid y = y) = \prod_{i=1}^d P(x_i \mid y = y)$$

El clasificador computa la probabilidad posterior para cada clase y como

$$P(y \mid \mathbb{X}) = \frac{P(y) \prod_{i=1}^d P(x_i \mid y)}{P(\mathbb{X})} \Rightarrow P(y) \prod_{i=1}^d P(x_i \mid y)$$

Nota Puede ignorarse $P(\mathbb{X})$ debido a que es un termino constante. Para esto se realiza una normalización con una constante ϵ de forma que $\sum_{y \in \mathbb{Y}} \epsilon^{-1} P(y \mid \mathbb{X}) = 1$.

Clasificación con SUPPORT VECTOR MACHINES (SVM)

Técnica de clasificación con una frontera de decisión en forma de hiper-planos que permiten aplicaciones con vectores de alta dimensionalidad. En su forma no-lineal se utiliza una función polinomial de similitud K para no afectar el rendimiento por la alta dimensionalidad.

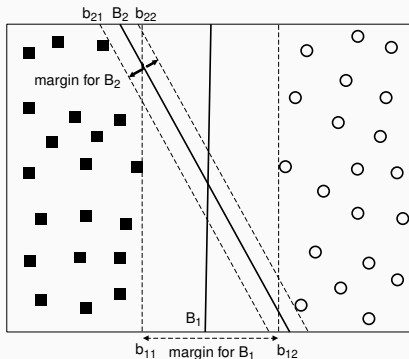


Figura 1: Maximum Margin Hyperplanes. Tomado de [3].

SELF-ORGANIZING MAPS (SOM)

Es un mapa discreto de o neuronas con vectores $\mathbf{w} \in \mathbb{R}^m$ que se adaptan a una entrada de $\mathbf{X} \in \mathbb{R}^{m \times N}$ de N patrones. Tiene una adaptación con una tasa de aprendizaje α_t y un área de afectación σ_t que se reducen por cada iteración $t \in \{0, \dots, T\}$.

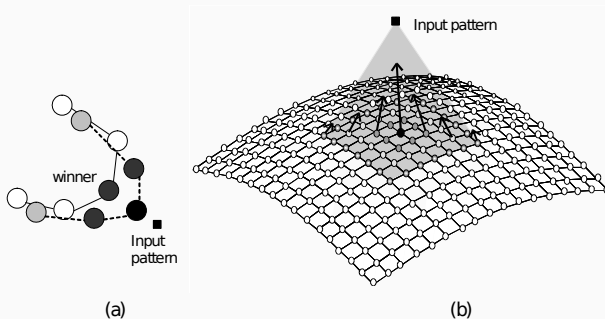


Figura 2: Proceso de adaptación de SOM, (a) uni-dimensional, (b) bi-dimensional. Tomado de [1].

Ejemplo de SOM

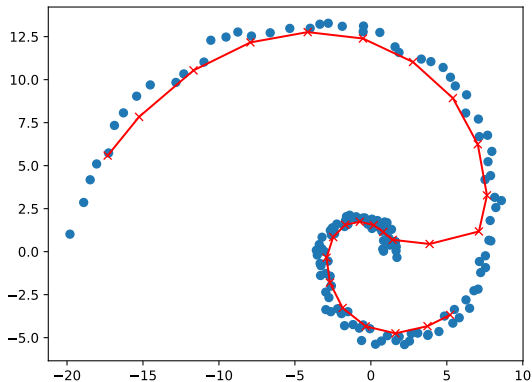


Figura 3: Ejemplo de salida de SOM uni-dimensional con 25 neuronas. Implementación propia.

Aplicación de SOM en perfilamiento de criminales

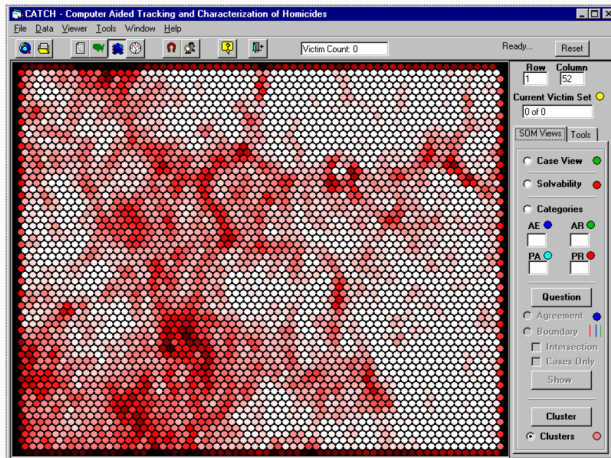


Figura 4: Ejemplo de uso de SOM en aplicaciones de perfilado. Tomado de [2].

Problemas y soluciones

Representación de palabras: BAG OF WORDS

N es el tamaño del diccionario de términos D (e.g. $N = |D|$).

$$\text{word2idx} = \left\{ (t_i, i) : \forall i \in \{1, \dots, N\} \right\}$$

$$\text{idx2word} = [t_1, \dots, t_N]$$

Representación de palabras en vectores para BoW

Para un término individual su vector representativo se define como:

$$\mathbf{e}^{(i)}, (t, i) \in \text{word2idx}$$

Para un documento d de términos, se calcula por cada término que existen dentro del diccionario su vector representativo como:

$$\mathbf{s} = \sum_{(t,i) \in \text{word2idx}} \mathbf{e}^{(i)}, t \in d$$

TF-IDF = Term Frequency – Inverse Document Frequency

$\text{tf}(t, d)$ = Frecuencia del termino (o n-grama) t en el documento d

$$\text{idf}(t, D) = \log\left(\frac{N}{|\{d \in D : t \in d\}|}\right); N = |D|$$

$$\text{tf-idf}(t, d, D) = \text{tf}(t, d) \cdot \text{idf}(t, D)$$

Regresión logística $\sigma(x)$ y One vs Rest

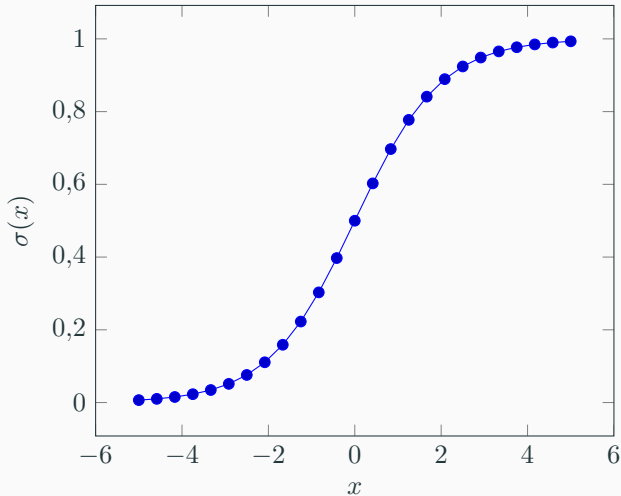


Figura 5: Gráfica de función sigmoide.

MODELO 1: Predicción de etiquetas de Twitter

Por hacer

MODELO 2: Reconocimiento de NAMED ENTITIES con redes LSTM

Por hacer

Conclusiones y trabajo futuro

Por hacer

Por hacer

- [1] L. N. De Castro.
Fundamentals of natural computing: basic concepts, algorithms, and applications.
Chapman and Hall/CRC, 2006.
- [2] J. Mena.
Investigative Data Mining for Security and Criminal Detection.
Elsevier Science, 2003.
- [3] P.-N. Tan, M. Steinbach, and V. Kumar.
Introduction to Data Mining.
Addison Wesley, us ed edition, May 2005.